心理学报 2023, Vol. 55, No. 7, 1–16
Acta Psychologica Sinica

https://doi.org/10.3724/SP.J.1041.2023.001

认知诊断测评中缺失数据的处理：随机森林阈值插补法*

游晓锋[1] 杨建芹[1] 秦春影[1] 刘红云[2,3]

([1]南昌师范学院数学与信息科学学院，南昌 330032)
([2]应用实验心理北京市重点实验室；3 北京师范大学心理学部，北京 100875)

摘　要　认知诊断测评中缺失数据的处理是理论和实际应用者非常关注的研究主题。借鉴随机森林插补法(RFI)不依赖于缺失机制假设的特点，对已有的 RFI 方法进行改进，提出采用个人拟合指标(RCI)确定插补阈值的新方法：随机森林阈值插补方法(RFTI)。模拟研究表明，RFTI 在插补正确率上明显高于 RFI 方法；与 RFI 和 EM 方法相比，RFTI 在被试属性模式判准率和边际判准率上表现出明显优势，尤其是非随机缺失和混合缺失机制，以及缺失比例较高的条件下，其优势更加明显。但对项目参数的估计，RFTI 方法与 EM 方法相比不具有优势。

关键词　缺失数据，认知诊断测评，随机森林阈值插补，随机森林插补，EM 算法

分类号　B841

1　引言

近年来，教育与心理评估的实践越来越关注测评结果的应用，随着信息技术的发展和精准测评服务的需求，测评日益融入日常的教学和学习过程(Bennett, 2010)。认知诊断测评(cognitive diagnosis assessment, CDA)通过被试在测验上的反应模式对其特定的知识结构(knowledge structure)和加工技能(processing skills)进行评价，而推知被试的知识状态(knowledge state, KS)，从而对其优势和劣势提供更具诊断性的信息。认知诊断测评由于其在测评结果反馈上的优势备受研究者和实践应用者的青睐，然而，实际测验中往往不可避免存在缺失数据。造成数据缺失的原因有多种，一方面测验设计上可能带来作答数据缺失，例如，国际大规模 PISA 测试、分层教学等个性化学习的测试，每个学生只完成全部测试的部分题目；另一方面，除设计造成的缺失数据外，由于其他原因产生的缺失数据也很常见，例如，由于测验时间限制或测验安全方面的考虑，以及测试者有意忽略测验中某些题目等(Cheema, 2014; Mislevy & Wu, 1988; Pohl et al., 2014; Rose et al., 2010)。大量的研究证实不同缺失值处理方法会对个体知识状态的估计精度带来不同影响(Dai, 2017; Pan & Zhan, 2020)。因此，在实际 CDA 测验中应重视缺失数据问题，并选用合适方法处理，以提升诊断精度(宋枝璘 等, 2022)。

根据以往的研究，基于认知诊断模型(Cognitive Diagnosis Model, CDM)的缺失数据的处理方法，大多借鉴项目反应理论(Item Response Theory, IRT)模型中处理缺失数据的方法。可以概括为以下三种：(1)传统的缺失值删除、单一插补或替换方法，删除方法主要包括列删除(Listwise)和对删除(Pairwise)，比较简单的替换方法是将缺失数据直接替换为 0，即零替换方法；(2)基于模型的处理方法，其基本思想是在模型参数估计的过程中通过似然函数处理缺失数据，其中典型的方法有期望

收稿日期：2022-04-23
* 江西省教育厅科技重点项目(GJJ212601)；南昌市教育大数据智能技术重点实验室(2020-NCZDSY-012)；国家自然科学基金项目(32071091)。
通信作者：刘红云，E-mail: hyliu@bnu.edu.cn

最大化算法(Expectation-Maximization algorithm, EM)和全息极大似然估计方法(Full Information Maximum Likelihood, FIML); (3)基于随机分布假设的多重插补方法，该方法的基本思想是基于假设的随机分布对缺失数据进行多次随机插补，其中典型的多重插补方法包括基于回归预测值分布的多重插补。研究者结合不同模型，对不同方法的表现进行了比较。Finch (2008) 结合 IRT 模型，对不同的缺失数据处理方法进行了比较，结果发现，很难找到一种方法，其表现在任何情况下均优于其他方法，不同的方法在不同的缺失机制下有各自的优缺点。Dai (2017) 首次结合认知诊断 DINA 模型，探讨了零替换、个体均值插补法、两步插补法、反应函数法(Response Function Imputation)和 EM 算法 5 种缺失数据处理方法在不同缺失比例和缺失机制条件下对项目参数和个体掌握模式的影响。研究发现，在 CDM 中，如果缺失数据被忽略或处理不当，则会对学生的属性掌握模式和项目参数的估计带来偏差；相比其它 4 种方法, EM 算法得到的个体属性掌握模式的判准率最高，且随着缺失比例增加, EM 算法的优势更加明显；对于项目参数的估计精度，则没有哪种方法在任何情况下都优于其他方法；总体而言，替换为零的方法和个体均值插补法对于 CDM 中的缺失数据处理不是好的选择, EM 算法相对表现最优。Dai 和 Svetina Valdivia (2022)结合 DINA 模型，比较了包含 FIML 和 EM 算法在内的 10 种缺失数据处理方法的表现，结果发现 FIML 和 EM 算法表现类似。Pan 和 Zhan (2020)在随机缺失机制的假设下结合追踪 CDM，探讨了缺失比例和测验长度的影响，研究发现缺失比例是影响参数估计和诊断结果精度的最主要因素，缺失比例超过 20%，诊断结果的精度就会明显变差，针对缺失比例较高的情况(不超过 40%)，可以通过增加测验长度弥补缺失数据带来的不利影响。宋枝璘等(2022)结合 GDINA 模型，在完全随机缺失、随机缺失和非随机缺失的条件下，比较了零替换、多重插补(Multiple Imputation, MI)、EM 算法和 FIML 方法的差异，结果发现在估计个体知识状态时，EM 算法和 FIML 表现较好，其中 EM 表现更优。在本研究中，我们将选择表现较好且稳定的 EM 算法与新提出的方法进行比较。

认知诊断模型中缺失数据的处理可以直接借鉴 IRT 模型中缺失数据的处理方法，其研究结论也与基于 IRT 模型缺失数据的处理方法一致，即相比于传统的缺失数据处理方法，基于模型的方法更有优势 (Schafer & Graham, 2002)。但是，这类方法仍然面临以下三方面的问题: (1)已有的缺失数据处理方法大多是基于完全随机缺失(missing completely at random, MCAR)和随机缺失(missing at random, MAR)机制假设的参数插补的方法，假设条件和模型限定较多，不能有效处理非随机缺失(missing not at random, MNAR)或混合(MIXED)机制下(De Ayala et al., 2001)的缺失类型(关于缺失机制的介绍可参考 Little 和 Rubin(2002)，宋枝璘等(2022)或本研究模拟设计部分的相关内容)。而在实践中，学生可能会因各种原因漏答部分试题，缺失数据产生原因具有很高的不确定性和复杂性，对于缺失机制的判别没有明确的衡量标准(De Ayala et al., 2001)。探索适用于不同缺失机制的缺失数据处理方法是目前尚未很好解决的问题之一。(2)已有的缺失数据处理方法无法有效处理缺失比例较高(>30%)的情况。纵观以往对缺失数据处理方法的模拟研究，设定的缺失比例从 2% (De Ayala et al., 2001)到 50% (Glas & Pimentel, 2008)不等，但大部分在 5%到 30%之间(Finch, 2008)。已有的缺失数据处理方法在缺失比例低(小于 20%)时表现良好，但在缺失比例超过 20%时应用效果已不明显。缺失比例超过 30%以上的研究不多见，且发现各种方法的估计偏差均较大。然而在实践中一些测验设计导致的数据缺失比例较高的情况并不少见，例如大规模测试中常用的分块设计(fractional block design) (McArdle, 1994)，平衡非完全分块螺旋设计(balanced incomplete blocks (BIB) spiral design) (Johnson, 1992)等常用的矩阵抽样设计的方法。这些设计中缺失数据的比例往往超过了 50% (Graham et al., 2006)。因此，发展能够较好处理高比例缺失数据的方法也是亟待解决的问题之一。(3)随着认知诊断理论在测评中的应用，以及近年来教学设计的改变和个性化学习的发展(如走班制和分层教学)，即便是日常的测评也不再是所有的学生同步完成相同的练习或测试，而是对不同的学生进行有区别的评估；同时，为了提高学习效率，教育测评实践也面临着如何基于更少的题目，得到较为精准的诊断结果的问题。因此，如何在不增加测验题目的情况下，提高缺失数据处理方法的精度也是认知诊断测评的实践需要。

近年来，随着教育测评理论和人工智能技术的不断发展，IRT、CDM 与机器学习相关技术的结合应用越来越受到国内外研究者的关注(Chen et al.,

2018; Zhang & Chang, 2016; 王璞珏，刘红云,2019)。机器学习的兴起也为缺失数据的处理带来了一种新的思维范式，即将数据集中的缺失值作为机器学习模型中的未知变量，将数据集中的具有特定变量观测值的数据样本用作机器学习模型的训练集，再将具有缺失值的数据样本输入训练后的模型，从而对缺失值进行插补(Liu & Gopalakrishnan, 2017)。已有研究发现，许多机器学习技术非常适合处理缺失数据的问题，并且在处理效果上比经典的统计处理技术表现得更好(Mabrey, 2006)。Stekhoven 和 Bühlmann(2012)在随机森林算法的基础上提出了随机森林插补(Random Forest Imputation, RFI)法，该方法是针对缺失数据处理提出的一种新的非参数插补方法。RFI 方法突出的特点是能处理不同类型的数据，能够利用所有可观察到的数据，并且对数据分布的假定前提条件很少。由于 RFI 方法的准确性和稳健性等诸多优点，它已在一些复杂研究中得到了应用(沈琳 等，2014)。然而，这一方法与测量模型的结合应用尚属空白，其方法的适用性和效果尚待检验。

综上，本文结合目前 CDA 实际中缺失数据处理遇到的困难，基于 RFI 的基本思想，提出一种在 CDM 模型下处理缺失数据的新方法：随机森林阈值插补(Random Forest Threshold Imputation, RFTI)方法。并通过两个 Monte Carlo 模拟研究验证新方法的表现。模拟研究一比较不同缺失机制和不同缺失比例条件下，RFTI 相对 RFI 方法对缺失数据插补正确率方面的优势，以验证所提出动态阈值方法的必要性；模拟研究二从个体属性模式判准率和项目参数估计精度两方面探讨 RFTI 方法的表现，并与传统表现较好的 EM 方法和 RFI 方法比较，探讨 RFTI 的优势和适用条件。

2　随机森林阈值插补法(RFT)的提出

本文的主要目的是提出一种新的适合于 CDM 缺失数据的处理方法，即随机森林阈值插补法。在介绍新的方法之前，我们首先介绍本文所使用的认知诊断模型，其次介绍已有的 RFI 方法，并对其局限性进行分析；最后在 RFI 的基础上提出改进阈值的随机森林阈值插补法，即 RFTI。

2.1　认知诊断模型：DINA 模型

DINA (Deterministic Inputs, Noisy “And” Gate Model, DINA)模型是一种非补偿的认知诊断模型，其中“非补偿”是指属性之间不具有互补性，被试只有完全掌握项目所需的所有属性才能正确答对该项目。由于 DINA 模型本身的定义简单，每个参数对应的含义具有可解释性的特点，近年来被广泛应用于认知诊断模型相关的理论研究和实际应用研究。例如：关于项目属性辅助标定(汪文义，2012)，含认知诊断功能的计算机化自适应测验的项目增补(陈平,辛涛,2011)，错误定义的 Q 矩阵下被试分类准确性(喻晓锋 等，2014)，以及与其他分类模型结合的问题(罗照盛 等，2015)，这些新方法的探索均是基于 DINA 模型开展的拓展研究。本研究我们也将基于 DINA 模型探讨不同缺失数据处理方法的表现，下面，首先简要介绍常用的 DINA 模型。

以 0-1 计分的题目为例, DINA 模型的定义如下：

$$P_j(\alpha_i) = P(X_{ij} = 1 \mid \boldsymbol{\alpha}_i) = g_j^{1-\eta_{ij}}(1-s_j)^{\eta_{ij}} = \begin{cases} g_j & 若\ \eta_{ij} = 0 \\ 1-s_j & 若\ \eta_{ij} = 1 \end{cases} \quad (1)$$

其中，s_j 为第 j 个题目对应的失误参数(slipping parameter)，其值介于 0 和 1 之间，用来描述被试在题目 j 上失误的概率，即被试在完全掌握了题目 j 所需属性的条件下，但是没有正确回答该题目的条件概率。g_j 为第 j 个题目对应的猜测参数(guessing parameter)，其值也介于 0 和 1 之间，与 s_j 参数相反，描述的是被试猜对第 j 个项目的概率，即被试没有完全掌握该项目考核的所有属性的条件下，但答对了该项目的条件概率。$\boldsymbol{\alpha}_i$ 表示知识掌握状态向量或属性掌握模式向量，其中的元素 $\boldsymbol{\alpha}_{ik}$ 表示被试 i 是否掌握属性 k，如果掌握 $\boldsymbol{\alpha}_{ik} = 1$，否则 $\boldsymbol{\alpha}_{ik} = 0$。如果被试 i 的属性掌握模式为 $\boldsymbol{\alpha}_i$，其在题目 j 上的理想反应模式可以表示为：

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} \quad (2)$$

其中，q_{jk} 为 Q 矩阵中对应题目 j 和属性 k 的值，如果题目 j 考核了属性 k，则 $q_{jk} = 1$，否则 $q_{jk} = 0$；Q 矩阵为描述测验题目与属性之间关系的矩阵(丁树良 等, 2012)。

2.2　随机森林插补法

RFI 是由 Stekhoven 和 Bühlmann (2012)提出的一种新的非参数插补方法(也称 missForest 算法)，该方法的基本思想和步骤如下。

假设一个 $n \times m$ 的数据集，n 表示被试的个数，m 为变量的个数，即测验中包含的题目数。用 $X = (X_1, X_2, \cdots X_m)$ 表示被试的作答数据集，X_h 为任意一个可能存在缺失值的变量，$i_{mis}^{(h)} \in \{1, 2, \cdots n\}$

为 X_h 上包含缺失值的被试集合。可以将其分成 4 部分:

(1)用 $y_{obs}^{(h)}$ 表示 X_h 的观测值, 即除所有 $i_{mis}^{(h)}$ 之外的被试在 X_h 上的观测数据;

(2)用 $y_{mis}^{(h)}$ 表示 X_h 的缺失值, 即所有 $i_{mis}^{(h)}$ 在 X_h 上的未观测到的数据;

(3)用 $x_{obs}^{(h)}$ 表示 X_h 上无缺失的被试在 X_h 以外所有变量上的数据, 即除所有 $i_{mis}^{(h)}$ 之外的被试在除 X_h 以外的其他所有变量上的数据(可能含缺失, 因为 $i_{mis}^{(h)}$ 只是在变量 X_h 上没有观测值);

(4)用 $x_{mis}^{(h)}$ 表示 X_h 上有缺失的被试在 X_h 以外所有变量上的数据, 即所有 $i_{mis}^{(h)}$ 在 X_h 之外的其他所有变量上的数据(也可能不含缺失)。

采用随机森林对缺失数据进行插补时, 将变量 X_h 上没有缺失的个体数据, 即 $y_{\text{obs}}^{(h)}$ 和 $x_{\text{obs}}^{(h)}$, 作为随机森林的训练样本集, 得到预测模型; 再基于预测模型对缺失数据 $y_{\text{mis}}^{(h)}$ 进行插补。具体的插补步骤如下:

首先, 采用传统的缺失数据插补方法, 如均数插补法计算 X 中所有缺失值的初值, 然后按照缺失值的数量升序将所有含缺失的变量 X_h 进行排序, 得到的结果的矩阵记为 $X_{(start)}^{imp}$,将其赋值到矩阵 X_{old}^{imp} 。

其次, 对于每一个变量 X_h, 使用随机森林算法对缺失数据进行插补, 分为如下两步: 第一步, 用因变量 $y_{\text{obs}}^{(h)}$ 和自变量 $x_{\text{obs}}^{(h)}$ 训练出一个 $y \sim x$ 的随机森林模型; 第二步, 将 $x_{\text{mis}}^{(h)}$ 作为特征变量输入, 用训练出的随机森林模型预测缺失值 $y_{\text{mis}}^{(h)}$ 。对所有 X_h 预测插补完成后, 所得到的矩阵记为 $X_{\text{new}}^{\text{imp}}$ 。

定义收敛指标, 这里的收敛指标是指迭代中插补值变化的情况小于某个标准 γ, 对于离散型变量, 表示插补值变化的指标的计算公式为:

$$\Delta_F = \frac{\sum_{j \in F} \sum_{i=1}^{n} I_{X_{new}^{imp} \neq X_{old}^{imp}}}{\# NA} \tag{3}$$

其中, F 表示两次迭代中被插补数据的集合, I 为指标变量, 记录了插补矩阵 X_{new}^{imp} 与 X_{old}^{imp} 相比第 i 行第 j 列的值是否发生变化, 若变化则记为 1, 否则记为 0, 因此上式中的分子表示两次迭代之间发生变化的插补值个数, $\# NA$ 是在离散变量中总的缺失值数量(Stekhoven & Bühlmann, 2012), Δ_F 描述了前后两次插补值变化的个数占总缺失数据个数的比例, 值越小表示两次迭代得到插补数据集的差异越小。

判断 X_{new}^{imp} 与 X_{old}^{imp} 的差异 Δ_F 是否满足迭代停止标准 γ, 如果 Δ_F 的值满足迭代停止标准 γ 则将 X_{new}^{imp} 作为最终的插补结果; 否则, 用本次插补得到的矩阵 X_{new}^{imp} 替换 X_{old}^{imp}, 不断重复迭代上述插补过程, 直到满足迭代停止标准或达到最大允许迭代次数。本研究中, 我们参考 Stekhoven (2013)的研究, 将 γ 值设定为 0.05, 最大允许迭代次数设定为 100 次。

对于 0-1 计分的题目, RFI 在对缺失数据进行插补时, 首先对于每一个缺失的未观测值, 得到一个处在[0,1]区间内的概率值, 用于表示该缺失数据取值为 1 的概率。该概率值越接近于 1 表示当前未观测值为 1 的概率越大, 该概率值越接近于 0 则表示当前缺失数据取值为 0 的概率越大。以 0-1 计分的题目为例, 实际应用中通过一个选定的概率阈值, 将所得概率值转换为 0-1 的二分值。RFI 方法一般将概率阈值设置为 0.5, 即当计算出的概率值大于 0.5 时, 将缺失值替换为 1; 当概率值小于 0.5 时, 将缺失值替换为 0; 当概率值正好等于 0.5 的时候, 缺失值随机替换为 0 或 1。不难理解, 不考虑教育测评的实际背景, 在概率值为 0.5 左右时, 无论将缺失值替换为 0 或者 1, 缺失值被错误插补的概率都非常大, 因为此时模型所做预测的不确定性非常大。如果结合教育测评实际, 当概率值在 0.5 或以下时, 认为其作答错误(即替换为 0)是比较合理的; 然而, 如果将概率值在 0.5 以上的都插补为 1, 则对于 0-1 计分的题目就过于宽松了。另外, RFI 固定阈值的方法没有考虑缺失数据插补过程中, 由于插补不准确性所带来的模型与数据拟合假设被违背的问题。鉴于此, 我们提出修正的动态阈值的随机森林插补方法。

2.3 随机森林阈值插补方法

2.3.1 认知诊断模型的个人拟合指数与阈值选择

本文提出的随机森林阈值插补方法的基本思想为, 在随机森林插补法的基础上设定两个概率阈值, 其中将概率转换为 0 的第一个阈值仍然采用 0.5, 而将概率转换为 1 的第二个阈值采用结合模型拟合指标的动态阈值。我们将个人拟合指数应用于动态阈值的确定。

Cui 和 Li (2015)将被试理想反应与观察反应之间的关系作为认知诊断模型的个人拟合的指标, 提出采用反应一致性指标(the response conformity index, RCI)描述学生的观察反应与基于 Q 矩阵得出的期望理想反应之间的一致性。RCI 的具体计算方法如下:

$$RCI_i = \sum_{j=1}^{m} \left| ln\left[-\frac{X_{ij} - P_j(\boldsymbol{\alpha_i})}{I_j(\boldsymbol{\alpha_i}) - P_j(\boldsymbol{\alpha_i})} \right]^{X_{ij}+I_j(\alpha_i)} \right| \tag{4}$$

其中，m 为试题的个数，X_{ij} 为被试 i 在第 j 道试题上的真实作答 $(i=1,2,\cdots,n; j=1,2,\cdots,m)$，$\boldsymbol{\alpha_i}$ 表示被试 i 的属性掌握模式，$P_j(\boldsymbol{\alpha_i})$ 表示属性掌握模式为 $\boldsymbol{\alpha_i}$ 的被试正确作答试题 j 的概率，$I_j(\boldsymbol{\alpha_i})$ 为属性掌握模式为 $\boldsymbol{\alpha_i}$ 的被试在试题 j 上的理想作答反应，对于 DINA 模型，即为 η_{ij}。$I_j(\boldsymbol{\alpha_i})$ 的值等于 0 或者 1，完全依赖于 $\boldsymbol{\alpha_i}$ 是否完全包含被试作答试题 j 所需要的所有属性，作答试题 j 所需要的属性由 Q 矩阵决定。当 $\boldsymbol{\alpha_i}$ 包含作答试题 j 所需要的所有属性时，$I_j(\boldsymbol{\alpha_i})$ 的值为 1；当 $\boldsymbol{\alpha_i}$ 不包含作答试题 j 所需要的所有属性时，$I_j(\boldsymbol{\alpha_i})$ 的值为 0。在实际的情况下，被试的真实掌握模式 $\boldsymbol{\alpha_i}$ 无法知晓，因此，在计算的时候采用估计得到被试的属性掌握模式 $\hat{\boldsymbol{\alpha}}_i$。

由公式(4)可以看出，RCI 指标描述的是被试的实际作答反应 X_{ij} 与被试的理想反应 $I_j(\alpha_i)$ 之间的偏离程度。被试的观察反应与理想反应差异越大，RCI 的值越大。因此，从理论上讲，缺失数据插补的正确率越高，观察反应与理想反应越一致，RCI 的值就越小；而缺失数据正确率越低，RCI 的值就越大。因此可以将 RCI 拟合指标作为不同阈值的评价标准以及对缺失数据插补的终止标准。即通过重复计算不同阈值插补数据集所对应的个人拟合指数，将个人拟合统计量的最优值所对应的插补作答矩阵作为最终的插补结果，与之对应的阈值即为重复过程中最终确定的第二个阈值。

由于 RCI 指标适用于所有明确定义项目反应函数 $P_j(\boldsymbol{\alpha_i})$ 的认知诊断模型，例如 GDINA，RUM 等模型(Cui & Li, 2015)。因此，上述提出的采用个人拟合指数 RCI 确定动态阈值的思想也同样适用于所有明确定义项目反应函数的认知诊断模型。下面为了叙述清晰，以 DINA 模型为例介绍随机森林阈值插补法的步骤。

2.3.2　随机森林阈值插补法的步骤

首先，采用 RFI 方法得到每一个缺失数据插补的概率值。第一概率阈值为 0.5，设第二概率阈值为 τ (用户设定 0.5 到 1 之间的数值)。当 RFI 方法得到的插补概率值 p 大于等于 τ 时，对应缺失值位置插入数值 1；当概率值 p 小于等于 0.5 的时候，对应的缺失值位置插入数值为 0；当概率值 p 小于 τ 并且大于 0.5 的时候，对应的缺失值位置保留缺失，不作插补。具体如下所示：

$$X_{ij} = \begin{cases} 1 & p_{ij} \geqslant \tau \\ NA & 0.5 < p_{ij} < \tau \\ 0 & p_{ij} \leqslant 0.5 \end{cases} \tag{5}$$

其中，X_{ij} 为第 i 个被试在第 j 道试题上的插补结果，p_{ij} 为第 i 个被试在第 j 道试题上 RFI 的概率值，NA 表示缺失，τ 为概率阈值。通过公式(5)的计算之后可以得到对应概率阈值 τ 的作答矩阵。

然后，采用 DINA 模型估计得到每个被试的属性掌握模式，计算相应的 RCI 个人拟合指数。由于插补后的作答矩阵 X 仍然有少量缺失，为了消除缺失数据的影响，对 RCI 指数计算公式进行校正，并对其求均值，具体校正后的公式如下：

$$mean_RCI_i = \frac{\sum_{j=1}^{m_i} \left| ln\left[-\frac{X_{ij} - P_j(\alpha_i)}{I_j(\alpha_i) - P_j(\alpha_i)} \right]^{X_{ij}+I_j(\alpha_i)} \right|}{m_i} \tag{6}$$

其中 m_i 为第 i 个被试插补后非缺失数据的试题个数，m_i 的取值介于 0 和 m 之间，即 $0 < m_i \leqslant m$。之后对所有被试求得 $mean_RCI_i$ 的均值，即为当前概率阈值 τ 所对应的拟合评价指标。在缺失值插补的过程中，对于 τ 取不同的值，根据公式(6)计算相应的 $mean_RCI_i$ 的均值，最后选取使得 $mean_RCI_i$ 的均值最小的 τ 作为最终缺失数据插补的第二个概率阈值。

具体实现过程中，阈值 τ 的取值在 0.5-1 之间，可按照步长 δ(研究者自定义，如 $\delta = 0.01$)递增，可以得到 T 个不同的阈值(例如，当 $\delta = 0.01$ 时，$T = 50$)。根据不同阈值 $\tau^{(t)} (t = 1,2,\cdots,T)$ 插补，得到 T 个插补后的作答矩阵 $X^{(t)} (t = 1,2,\cdots,T)$。基于 $X^{(t)}$，根据公式(6)可以计算得到被试 i 的个人拟合指数 $mean_RCI_i^{(t)}$，计算其均值 $mean_RCI^{(t)} = \sum_i mean_RCI_i^{(t)} / n$。在 K 个拟合指标中选取最小的 $mean_RCI_{min} = min(mean_RCI^{(t)})$，将 $mean_RCI_{min}$ 对应的阈值作为最后确定的阈值，其对应的插补后的作答矩阵 X 作为最终的插补后的作答数据。实际中，根据插补的目的我们只需要最后的插补数据集即可。

2.3.3　随机森林阈值插补法的算法实现

采用 R 语言来实现随机森林阈值插补方法，其具体的算法步骤如下：

(1)导入带有缺失值的作答矩阵数据集，设为 missData；

(2)设置概率阈值 τ 的范围和步长，τ 值的默认范围为 0.5-1，默认步长为 0.01；

(3)将 missData 数据集赋值给 oldData;

(4)采用 missForest 算法对 oldData 中的缺失值进行插补，得到带有插补概率值的矩阵 imputeData;

(5)根据概率阈值 τ 对 imputeData 作答矩阵重新赋值，将其转换为新的作答矩阵数据集 newData。

(6)比较 oldData 与 newData，如果 oldData 和 newData 两个作答矩阵不完全相同，则将 newData 数据集赋值给 oldData，重复第(4)(5)步，直到 oldData 与 newData 完全相同，迭代终止;

(7)对最终的 newData 数据集，采用 EM 方法估计 DINA 模型的项目参数，采用 MAP 方法估计被试的掌握模式，并计算出所有被试个人拟合指标 mean_RCI 的均值;

(8)重复(2)-(7)步，计算所有概率阈值 τ 对应的个人拟合指标 mean_RCI 的均值。

(9)找出最小的个人拟合指标 mean_RCI 均值所对应的最终作答矩阵 newData 数据集作为最终输出的插补结果数据集。

基于上述步骤，我们在 R 语言平台 R-3.5.3 版本上自编了 missForestDINA 函数，以方便研究者和应用者使用本研究提出的方法。使用时需事先安装 RFI 方法的 missForest 包(Stekhoven, 2013)和用于认知诊断模型参数估计和数据分析的 CDM 包(George et al., 2016)，然后调用 missForestDINA 的主函数。missForestDINA 主要函数为:

$missForestDINA(missData, Q, stepV = 0.01)$。

其中 missData 为输入的不完整数据集，Q 为认知诊断模型对应的 Q 矩阵，stepV 为迭代步长，函数返回的是插补后的数据集。missForestDINA 包的使用非常方便，使用者只需要读入含有缺失数据的文件和 CDM 的 Q 矩阵，并设置好搜索最小 RCI 值的迭代步长，就能得到插补到的数据集。

值得说明的是，以上随机森林阈值插补方法的步骤和算法实现虽以 DINA 模型为例，但是这一方法本身并不局限于 DINA 模型，对于明确定义了项目反应函数的认知诊断模型，只需将上述公式(6)中的 $P_j(\boldsymbol{\alpha}_i)$ 替换为对应的模型即可。这一方法不仅适用于大多数的认知诊断模型，而且适应于同一测验中的各个项目拟合模型可能不同的情况，例如，在同一个测验中，有些项目适合 DINA，而有些项目适合 A-CDM，有些项目则可能适合 GDINA 模型等(Liu et al., 2019; 刘彦楼 等, 2019)，此时，则需将对应项目的 $P_j(\boldsymbol{\alpha}_i)$ 替换为相应所适合认知诊断模型的项目反应函数。总而言之，基于认知诊断测评模型提出的随机森林阈值插补法具有较广泛的适用性。

3 研究 1: 随机森林阈值插补方法的插补率和正确率

从 RFTI 的步骤和公式(5)可以看出，该方法和其他的缺失数据处理方法不同，它是一种非完全的插补方法。虽然插补后的数据集仍可能包含一定比例的缺失数据，但是我们预期这一部分的比例应该较低，在后续分析中可以采用简单默认的缺失数据处理。另外，我们预期动态阈值的插补方法相对于 RFI 方法有较高的插补正确率。为了验证我们的预期，研究 1 的主要目的是，探讨在不同的缺失机制和缺失比例的情况下，RFTI 方法的插补正确率，以及插补后数据集的缺失比例，并将其插补结果直接与 RFI 法进行比较。

3.1 数据缺失机制

研究 1 考虑的主要影响因素为数据的缺失机制和缺失比例。

(1)数据缺失机制：考虑 MIXED、MNAR、MAR 和 MCAR 四种缺失数据的机制。

(2)缺失比例：本实验借鉴以往研究对缺失比例的设置条件(Dai, 2017)以及实际测验可能面临的情况，考虑 10%、20%、30%、40%、50%五种不同的缺失比例。

因此，共有 4×5=20 种实验条件，每种组合的条件下，重复模拟生成 100 个符合条件的被试作答数据集。再针对每个数据集分别使用 RFI 和 RFTI 两种缺失数据处理方法对缺失数据进行插补，主要借助 missForest 软件包和自编 missForestDINA 函数实现。

参考已有的认知诊断相关文献，其他条件设定如下。大多数研究设定的属性个数为 6 个 (Cheng, 2010; Gierl et al., 2011)，因此本实验的属性个数也设置为 6 个。de la Torre 等(2010)在研究样本量对 DINA 模型参数估计的影响时，使用的样本量是 500、1000、2000、4000 四个水平，结果发现，当样本量为 1000 时，DINA 模型能得到非常精确的参数估计。本研究将样本量设置为 1000，题目数设置为中等长度 30。假设属性间不存在层级关系，试题属性分配方式是随机的，但是每个属性至少存在 1 个测量单一属性的题目，每个属性至少有 3 个题目测量以保证模型可识别(Xu & Zhang, 2016)。

3.2 数据生成

3.2.1 参数生成

(1)被试知识状态

被试知识状态的生成依赖于多元离散均匀分布和多元标准正态阈值模型(Multivariate Normal Threshold Model)。对于多元标准正态阈值模型，先根据多元标准正态分布，模拟生成各被试在各属性上的随机数 $\boldsymbol{\theta} \sim \text{MVN}\left(0, \sum\right)$，其中协方差阵∑中主对角线元素全为 1，其他元素全设为 0.5。然后根据属性阈值对连续的随机数 θ_{ik} 进行 0-1 化，得到被试 i 对属性 k 的掌握状态 $\boldsymbol{\alpha}_{ik}$，公式如下：

$$\boldsymbol{\alpha}_{ik} = \begin{cases} 1 & \theta_{ik} \geqslant \Phi^{-1}\left(\dfrac{k}{K+1}\right) \\ 0 & otherwise \end{cases} \tag{7}$$

其中 $i = 1, 2, \cdots, n, k = 1, 2, \cdots, K, \Phi^{-1}$ 为标准多元正态分布的逆累积分布函数，表示累计概率为 $\dfrac{k}{K+1}$ 对应的向量。

(2)项目参数

DINA 模型中的猜测参数 g 和失误参数 s 均从均匀分布中抽取，取值区间为[0.05, 0.25]。

3.2.2　完整作答数据的生成

采用 Monte Carlo 模拟方法，首先生成被试的知识状态(掌握模式)真值 $\boldsymbol{\alpha}$，项目参数真值(s 和 g)以及 Q 矩阵。为了保证模型可识别，在设定 Q 矩阵时，我们根据 Xu 和 Zhang (2016)的研究，首先随机生成只测量 K 个属性中某个单一属性的 K 道项目，即在 Q 矩阵中存在一个 K×K 的单位矩阵；其次，对于其他项目所测量的属性，要求每个属性至少还有 2 个项目测量这一属性，但是同一项目可同时测量多个属性。然后根据 DINA 模型的项目反应函数模拟生成被试的作答数据集。具体来说，模拟生成被试 i 在项目 j 上的得分时，依据 DINA 的项目反应函数计算其正确作答概率 $P_j(\boldsymbol{\alpha}_i)$，然后产生服从 U(0, 1)分布的随机数 r，如果 $r > P_j(\boldsymbol{\alpha}_i)$，则被试 i 在项目 j 上的得分为 0，否则得分为 1。

3.2.3　缺失数据的生成

由于使用 RFI 或 RFTI 方法进行插补时，需要首先基于目标变量上未缺失被试的数据训练模型，因此，对于生成的完整作答数据，从中随机选取 80%的被试作答数据用于生成缺失数据，剩下 20%的被试保留完整数据集，作为随机森林的训练样本集。需要说明的是，实际中完整的训练数据集并非必须的(Stekhoven, 2013)。

(1) MCAR 缺失数据的生成

MCAR 缺失机制指的是数据的缺失是完全随机的，不依赖于任何变量，即不论其它变量(如题目难度、区分度、被试能力值等)如何变化，数据产生缺失的概率都是均等的。根据 MCAR 的定义，MCAR 数据的生成是一个完全随机的过程，当数据总体缺失比例确定以后，可以通过产生随机数的方式来确定被试及某一题目的缺失，缺失的产生并不依赖于被试的能力及项目本身。例如，当数据总体缺失比例被设置为 30%的时候，针对每个被试在每道试题上的作答都生成一个 0 和 1 之间的随机数 r 来判断当前作答是否被设置为缺失，当随机数 r 小于缺失比例 0.3 时，试题作答被设置为缺失。由 R 语言 missForest 包中的 prodNA 函数具体实现该过程。

(2) MAR 缺失数据的生成

MAR 缺失机制指的是数据缺失的概率不是随机的，会受到数据集中已观测到的其他变量的影响，但不受缺失数据自身的影响。根据 MAR 的定义，MAR 数据的生成借鉴 De Ayala 等人(2001)及 Peugh 和 Enders (2004)提出的方法。首先，计算除目标题目外，每个被试的正确作答题目个数；然后，依据被试的得分确定每个被试作答的缺失比例，得分越高的被试其缺失作答的比例越小，得分越低的被试其缺失作答的比例越高。具体而言，首先基于完整的模拟数据集计算每个被试在各项目上的 CTT 得分，然后将被试的得分进行正态化转换，通过正态累积分布函数找到百分等级位于 5%、15%、30%、70%、85%、95%位置上的百分位数，根据这些百分位数将被试分成 7 组，设定得分越高的组数据缺失比例越低。用 MR 表示总缺失比例，则每组被试对应的缺失比例如表 1 所示。例如，对于总缺失比例 MR 为 30%的条件，原始得分处在 0%~5%这一区间的被试，其缺失比例为 1.5 × 30% = 45%, 5%~15%这一区间的被试，其缺失作答比例为 1.35 × 30% = 40.5%，依次类推。在确定了各区间被试作答的缺失比例后，再针对每个被试在每道试题上的作答都

表 1　不同分数段 MAR 缺失比例分布

分类分段	缺失比例(%)
0%~5%	MR×1.50
5%~15%	MR×1.35
15%~30%	MR×1.15
30%~70%	MR×1.00
70%~85%	MR×0.85
85%~95%	MR×0.65
90%~100%	MR×0.50

生成一个 0 和 1 之间的随机数 r 来判断当前作答是否被设置为缺失，当随机数 r 小于缺失比例时，试题作答被设置为缺失。

(3) MNAR 缺失数据的生成

MNAR 缺失机制指的是数据缺失的概率与缺失变量本身相关。对于 MNAR 缺失数据的生成，本研究借鉴了 Dai (2017)的研究，根据被试在每道试题上的作答计算该作答的缺失概率。即，若某被试在某试题上的原始作答正确，则缺失概率小，并且试题越难缺失比例越大，反之亦然。具体过程如下：首先，根据数据缺失的整体比例计算出每个被试缺失试题的个数；然后，为每个被试设置一个概率调节因子 ε，初值为 0，计算被试正确作答某道试题的概率 p，再通过产生一个 0 和 1 之间的随机数 r 来判断被试在某试题上的作答是否被设定为缺失，当随机数 $r > p+\varepsilon$ 时，被试在该试题上的作答则被设置为缺失。过程中如果实际缺失个数与初始设定的个数不相等，则重新设置 ε 的值。若缺失个数大于预设的缺失个数，则增大 ε 的值；若缺失个数小于预设分配缺失个数，则减小 ε 的值。不断调整 ε 的值并重新生成缺失数据，直到各被试缺失作答的试题个数等于初始设定的个数时结束。

(4) MIXED 缺失数据的生成

混合缺失机制是指缺失数据集中包含两种或以上的缺失机制。本研究借鉴了 De Ayala 等人(2001)和 Dai (2017)及 Peugh 和 Enders (2004)提出的方法。首先，采用与生成 MAR 缺失数据时相同的方法，将被试分为 7 组，并计算各组被试缺失作答的比例，使得分越高的被试的数据缺失比例越低。然后，计算出每个被试的数据缺失个数后，再采用 MNAR 缺失数据产生的过程得到所有被试的缺失数据。这样可以使得 MIXED 缺失数据的生成不仅依赖于被试能力，而且依赖于测验项目本身的特征。

3.3　评价指标

本研究用来评价插补效果的指标主要有：(1)缺失数据插补的正确率，描述的是缺失数据插补正确的个数占插补数据个数的比例，数值越大表示插补越准确。在本研究中由于 RFI 和 RFTI 插补为 0 的数据个数相同，我们只统计插补为 1 的正确率，以考察动态阈值的效果。(2)插补后数据集中仍然缺失的数据所占比例，用来描述 RFTI 插补后仍然缺失的数据占总数据个数的比例，其数值越小表明插补率越高。如果其比例较低(20%以内)，则说明前面提出的采用模型默认的方法处理少量没有插补缺失数据是合理的。

3.4　研究结果

表 2 呈现了不同缺失机制和缺失比例下，采用 RFI 方法和 RFTI 方法插补值为 1 时的正确率结果。表 3 呈现了不同缺失机制和缺失比例下 RFTI 方法的正确率和插补后仍缺失的数据比例。

表 2　不同缺失机制和比例下，RFI 方法与 RFTI 方法的插补正确率比较

缺失比例	RFI 插补值为 1 的正确率(%)				RFTI 插补值为 1 的正确率(%)			
	MIXED	MNAR	MAR	MCAR	MIXED	MNAR	MAR	MCAR
10%	49.39	59.19	75.54	75.30	71.80	78.57	82.12	83.07
20%	42.84	49.29	73.23	73.62	67.25	75.45	83.04	81.81
30%	35.42	44.98	71.49	71.65	68.26	74.91	80.35	81.48
40%	32.51	42.97	68.32	69.04	58.22	71.59	79.74	79.84
50%	30.89	42.60	66.74	64.97	49.44	64.58	76.67	78.09
平均	38.21	47.80	71.06	70.92	62.99	73.02	80.39	80.86

注：表中数据为插补为 1 时的插补正确率。

从表 2 可以看出，所有条件下，采用 RFTI 方法的插补正确率都明显高于 RFI 方法。缺失机制是影响插补率的主要因素，在缺失机制为 MIXED 和 MNAR 时，对于各缺失比例平均正确率，RFTI 方法比 RFI 方法要高出约 25%。在缺失机制为 MCAR 和 MAR 时，RFI 方法的插补正确率也要低于 RFTI 方法大约 10%左右。另外，随着缺失比例增加，两种方法的插补正确率均出现下降的趋势，但是 RFI 方法下降更快。

表 3　不同缺失机制和比例下，RFTI 方法的插补正确率和插补后的缺失率(%)

缺失比例	MIXED		MNAR		MAR		MCAR	
	正确率	缺失率	正确率	缺失率	正确率	缺失率	正确率	缺失率
10%	86.15	0.96	84.69	1.16	77.68	0.94	77.94	1.01
20%	85.86	2.13	84.39	2.81	77.97	2.03	77.69	2.02
30%	85.86	3.87	84.35	5.88	78.19	3.55	78.28	3.50
40%	85.61	7.27	84.38	9.03	78.27	5.98	78.48	5.60
50%	85.03	10.12	82.61	11.66	78.28	7.03	78.41	7.98

注：缺失率是指采用 RFTI 方法插补后，数据集中没有被插补数据所占比例。

从表 3 可以看出在同一缺失机制下，数据正确率的变化受缺失比例的影响不明显。但不同机制下插补的正确率存在差异。当缺失机制为 MIXED 时，

不同缺失比例条件下的正确率都达到85%以上；当缺失机制为MNAR时，插补的正确率与MIXED机制下的结果类似；但是当缺失机制为 MAR 和 MCAR 时，插补的正确率均在 78%左右，略低于MIXED和MNAR机制下的结果。这一结果与随机森林方法本身的特点有关，由于MIXED和MNAR机制下，被试的缺失模式反而可以为 RFTI 方法的训练模型提供更多的关于缺失反应模式的信息。

表 3 缺失率的结果表明，采用 RFTI 方法对原始数据进行插补后，数据的缺失率随着缺失比例的增加呈现上升的趋势。当缺失比例为 10%时，4 种缺失机制下插补后的缺失率均在 1%左右；当缺失比例为 30%时，MIXED、MAR 和 MCAR 三种缺失机制下的插补后缺失率均在 3%左右，MNAR 机制下也仅为 4%左右。当缺失比例为 50%的时候，MIXED 和 MNAR 机制下，插补后的缺失率为 10%左右，而 MAR 和 MCAR 机制下的插补后缺失率略低一些，均不超过 8%。

从研究 1 的结果可以看出，对 RFI 方法进行改进后的RFTI方法对于插补值为1时的正确率的提高有明显效果，并且采用RFTI方法处理后的数据的缺失比例都在 10%左右，因此，对基于 RFTI 方法处理后的数据进行后续分析时，可以采用简单忽略方法。

4　研究 2：随机森林阈值插补方法的效果检验

研究2的主要目的是探讨不同缺失机制和缺失比例下，RFTI方法相比于其它常用的缺失数据插补方法的优势。验证 RFTI 方法在 DINA 模型下处理缺失数据的效果，并且与 EM 算法和 RFI 方法进行对比。同时探讨数据缺失机制和缺失比例以及不同缺失数据处理方法对被试属性模式判准率、属性边际判准率及项目参数估计精度的影响。

4.1　研究设计

本研究的设定条件与研究 1 相同。考虑与缺失相关的因素有两个：缺失机制(MIXED、MNAR、MAR、MCAR)和缺失比例(10%、20%、30%、40%、50%)。共有 4×5=20 种组合，在每一种被试间变量组合的条件下，重复模拟生成 100 个符合条件的被试作答数据集，每个数据集分别采用 EM、RFI 和 RFTI 三种缺失数据处理方法进行分析。其他条件与研究 1 的设定相同。

4.2　研究方法

模拟数据生成方法与研究 1 相同。对于每种方法插补后的数据集，采用 EM 算法估计 DINA 模型的项目参数，采用后验概率估计法(Maximum A Posteriori, MAP)估计被试属性掌握模式。对于 RFTI 方法中插补后数据集中的缺失数据，采用忽略缺失数据的方法进行处理，即在估计被试掌握模式时将这个被试缺失的题目删除，估计题目参数时将在这道题目上缺失的被试删除。

4.3　评价指标

关于项目参数的估计，本研究主要采用了 2 个评价指数，分别为所有题目偏差 Bias 和均方根误差 RMSE 的均值。所有项目参数估计的偏差均值定义为：

$$Bias = \sum_{j=1}^{m}\sum_{r=1}^{R}(\hat{\pi}_{rj} - \pi_{rj}) / (\mathrm{R} \times m) \tag{8}$$

所有题目上平均的均方根误差定义为：

$$RMSE = \frac{\sum_{j=1}^{m}\sqrt{\frac{\sum_{r=1}^{R}(\hat{\pi}_{rj} - \pi_{rj})^2}{R}}}{m} \tag{9}$$

其中，R 表示独立重复模拟的次数，本研究中 R=100；m 表示题目的个数，本研究中 m =30，π_{rj} 和 $\hat{\pi}_{rj}$ 分别表示第 r 次重复第 j 个题目参数的真值和估计值，项目参数指 DINA 模型的失误参数 s 和猜测参数 g。Bias 指标反映了估计值与真值的偏差的平均值。Bias 越接近 0 表示能力估计越准确。RMSE 指标反映了项目参数真值与估计值的偏移均方根，其值越小表示估计准确性越高。

关于被试的知识状态估计结果，本研究采用了被试属性模式判准率(Pattern Match Ratio, PMR)和被试属性边际判准率(Marginal Match Rate, MMR)两个评价指标。

若被试属性掌握模式的估计向量 $\hat{\boldsymbol{\alpha}}$ 与真值向量 $\boldsymbol{\alpha}$ 相等，即对应的元素完全相同，则认为被试属性掌握模式的估计结果是正确的，记为 1；否则认为是错误的，记为 0。模式判准率 PMR 为 R 次重复中 n 个被试中属性掌握模式判断正确的人数 PN 所占的比例的均值，考查的是对属性掌握模式整体的判断准确性，其计算公式为：

$$PMR = \sum_{r=1}^{R}\frac{PN_r}{n} / R \tag{10}$$

其中 PN_r 为第 r 次重复中属性掌握模式判断正确的人数。

边际判准率 MMR 考查模型在每个属性上的平均判断正确的效果。首先统计每次重复中各个属性

k 上判断正确的被试人数 PN_k 占总人数 n 的比例, $k=1,2,\cdots,K$ (K 为属性总数), 然后对各属性的边际判准率求平均, 得到所有属性的平均边际判准率, 即:

$$MMR=\sum_{k=1}^{K}\sum_{r=1}^{R}PN_{rk}/(K\times R) \tag{11}$$

其中 PN_{rk} 为第 r 次重复中属性 k 判断正确的人数。

公式(10)和(11)中, R 表示独立重复模拟的次数; n 表示被试的人数, K 为考查属性个数。PMR 和 MMR 越高, 表示对被试掌握模式判断的准确性就越高。

4.4 研究结果

4.4.1 不同方法被试知识状态估计结果的差异

不同缺失机制和缺失比例下, 被试属性模式判准率(PRM)和属性边际判准率(MMR)结果见表 4。从表 4 的结果可以看出, 无论在哪种条件下, RFTI 方法在 PRM 和 MMR 上的估计结果均优于 EM 和 RFI 方法。

表 4 不同缺失机制和缺失比例下各缺失数据处理方法所得模式判准率和边际判准率

缺失机制	缺失比例	模式判准率(PMR)			边际判准率(MMR)		
		EM	RFI	RFTI	EM	RFI	RFTI
MIXED	10%	0.498	0.518	0.526	0.827	0.838	0.842
	20%	0.471	0.502	0.529	0.816	0.835	0.846
	30%	0.414	0.457	0.513	0.791	0.819	0.843
	40%	0.399	0.414	0.525	0.784	0.807	0.848
	50%	0.335	0.346	0.489	0.753	0.777	0.836
MNAR	10%	0.517	0.541	0.547	0.840	0.851	0.854
	20%	0.499	0.537	0.562	0.830	0.850	0.859
	30%	0.427	0.478	0.546	0.802	0.832	0.855
	40%	0.392	0.434	0.543	0.792	0.823	0.859
	50%	0.316	0.364	0.489	0.755	0.792	0.841
MAR	10%	0.482	0.482	0.486	0.825	0.829	0.830
	20%	0.430	0.434	0.439	0.797	0.807	0.810
	30%	0.370	0.377	0.384	0.774	0.787	0.792
	40%	0.349	0.355	0.366	0.754	0.771	0.778
	50%	0.281	0.285	0.298	0.716	0.737	0.749
MCAR	10%	0.462	0.463	0.467	0.819	0.824	0.825
	20%	0.432	0.437	0.442	0.797	0.805	0.808
	30%	0.374	0.379	0.386	0.770	0.783	0.789
	40%	0.341	0.345	0.357	0.750	0.767	0.776
	50%	0.302	0.305	0.319	0.727	0.747	0.760

缺失机制对不同方法之间的差异有明显的影响, 无论缺失比例大小, MNAR 和 MIXED 缺失机制时, RFTI 方法的优势更明显。为了清楚的呈现这一趋势, 我们以缺失比例 30%为例说明三种不同方法在不同缺失机制上的差异(表 4)。从表 4 的结果可以看出, 在不同的缺失机制下, 采用 RFTI 方法时的 PMR 均高于其他方法, 特别是在缺失机制为 MIXED 和 MNAR 时优势更加明显。当缺失机制为 MAR 和 MCAR 时, RFTI 仍优于其他两种方法, 但是三种方法之间的差异不大。另外, 在 MMR 上, RFTI 方法也均略高于其它方法, MIXED 和 MNAR 缺失机制下, 优势略微明显。但整体来讲, 由于 MMR 整体较高, 方法之间的差异不明显。

缺失比例影响在不同缺失机制下也表现出近似一致的趋势, 无论何种缺失机制, RFTI 在 PMR 和 MMR 上的表现均最优, 而且这一优势随着缺失比例的增加优势越来越明显。从表 4 可以看出, 当缺失比例为 10%的时候, RFI 方法和 RFTI 方法间的差异不明显, 但均高于 EM 方法。随着缺失数据比例的增加, 三种方法的 PMR 都随之下降, 但 RFT 方法下降的幅度最小。从 MMR 的结果来看, RFTI 方法也优于其它两种方法, 方法间差异随着缺失比例增大而增大。

4.4.2 不同方法项目参数估计结果比较

不同缺失机制和缺失比例下, 采用 EM、RFI、RFTI 三种方法在 DINA 模型 s 参数和 g 参数上的估计偏差和均方根误差的结果分别见表 5 和表 6。

表 5 不同缺失机制和缺失比例下各处理方法参数估计偏差

缺失机制	缺失比例	s 参数			g 参数		
		EM	RFI	RFTI	EM	RFI	RFTI
MIXED	10%	0.005	0.008	0.009	0.005	–0.011	–0.015
	20%	0.012	0.012	0.012	0.014	–0.016	–0.027
	30%	0.026	0.028	0.022	0.022	–0.024	–0.041
	40%	0.040	0.043	0.023	0.033	–0.026	–0.051
	50%	0.060	0.060	0.033	0.045	–0.027	–0.061
MNAR	10%	0.010	0.015	0.017	0.003	–0.012	–0.016
	20%	0.022	0.025	0.028	0.011	–0.018	–0.028
	30%	0.051	0.054	0.044	0.016	–0.026	–0.042
	40%	0.073	0.067	0.051	0.022	–0.028	–0.051
	50%	0.091	0.080	0.067	0.026	–0.035	–0.060
MAR	10%	0.015	0.026	0.028	0.004	–0.009	–0.011
	20%	0.029	0.049	0.054	0.010	–0.016	–0.020
	30%	0.049	0.078	0.087	0.015	–0.023	–0.030
	40%	0.067	0.081	0.092	0.021	–0.029	–0.039
	50%	0.094	0.110	0.120	0.027	–0.035	–0.048
MCAR	10%	0.015	0.032	0.035	0.004	–0.010	–0.012
	20%	0.030	0.052	0.058	0.010	–0.016	–0.020
	30%	0.050	0.077	0.085	0.014	–0.024	–0.030
	40%	0.066	0.089	0.098	0.021	–0.031	–0.040
	50%	0.094	0.109	0.110	0.027	–0.034	–0.047

表 6　不同缺失机制和缺失比例下各处理方法参数估计均方根误差

缺失机制	缺失比例	s 参数			g 参数		
		EM	RFI	RFTI	EM	RFI	RFTI
MIXED	10%	0.038	0.040	0.040	0.019	0.021	0.021
	20%	0.040	0.044	0.040	0.026	0.033	0.032
	30%	0.056	0.067	0.061	0.034	0.048	0.045
	40%	0.064	0.087	0.060	0.045	0.060	0.056
	50%	0.084	0.123	0.076	0.059	0.083	0.068
MNAR	10%	0.059	0.063	0.064	0.018	0.022	0.022
	20%	0.048	0.064	0.062	0.024	0.033	0.032
	30%	0.074	0.110	0.086	0.030	0.046	0.047
	40%	0.100	0.141	0.099	0.036	0.058	0.056
	50%	0.131	0.180	0.126	0.043	0.072	0.067
MAR	10%	0.053	0.067	0.068	0.018	0.020	0.020
	20%	0.055	0.092	0.092	0.022	0.029	0.028
	30%	0.076	0.135	0.135	0.027	0.037	0.037
	40%	0.091	0.156	0.154	0.032	0.048	0.048
	50%	0.126	0.201	0.192	0.038	0.056	0.057
MCAR	10%	0.048	0.067	0.067	0.018	0.021	0.021
	20%	0.060	0.095	0.096	0.023	0.029	0.029
	30%	0.081	0.136	0.136	0.026	0.038	0.038
	40%	0.092	0.166	0.161	0.032	0.047	0.048
	50%	0.129	0.206	0.186	0.038	0.057	0.057

从表 5 的结果可以看出，随着缺失比例增大，3 种方法对项目参数的估计偏差均有增大的趋势。对于项目参数 s，在 4 种不同缺失机制下，无论采用何种缺失数据处理方法，s 的值都被高估。在缺失机制为 MIXED 和 MNAR 时，缺失比例较低时(≤20%)，三种方法之间差异较小，EM 算法表现出微弱优势，而缺失比例较高时(≥30%)采用 RFTI 处理方法得到的 s 的估计偏差最小，EM、RFI 方法表现相当，并且随着缺失比例增加 RFTI 方法的优势更为明显。当缺失机制为 MAR 和 MCAR 时，采用 EM 方法得到的 s 的估计偏差最小，采用 RFTI 方法得到的 s 估计偏差最高。对于项目参数 g，无论在哪种缺失机制下，采用 EM 方法时 g 的值存在高估现象，采用 RF 和 RFT 方法时 g 的值都被低估，但偏差均较 s 参数小。

从表 6 估计均方根误差的结果可以看出，对于项目参数 s 的均方根误差的估计精度，大部分条件下 EM 方法的表现要优于 RFI 和 RFTI 方法，只有在 MNAR 和 MIXED 机制下且缺失比例高时，RFTI 方法表现出优势。对于项目参数 g，采用 EM 方法在 4 种缺失机制下的表现都是最好，RFI 和 RFTI 方法则表现相当。

5　讨论与结论

5.1　讨论

本研究尝试将机器学习中随机森林缺失数据的插补(RFI)方法应用于认知诊断模型，基于 RFI 方法将缺失数据插补为 1 时的正确率偏低的问题，提出了一种基于认知诊断模型中的个人拟合指标 RCI 来动态确定阈值的新方法，即随机森林阈值插补方法(RFTI)。该方法首次实现了缺失数据插补过程中，机器学习方法与认知诊断模型的结合应用，正确率和插补率的结果证实了这是一种有效的动态选择阈值的方法。

为验证该方法有效改进了 RFI 方法插补正确率过低的问题，我们以 DINA 模型为例，探讨了不同缺失比例和不同机制下，RFTI 方法对缺失数据的插补效果，结果证实了我们的假设和预期，RFTI 方法对于插补值为 1 时的正确率相对于 RFI 方法有明显提高，并用在各种实验条件下数据的插补率和正确率都有较好的结果；从整体正确率来看，采用 RFTI 方法比 RFI 方法有显著提高。由于其第二阈值的选择过程中考虑到了错误插补可能带来的对认知诊断模型拟合的破坏，这一方法阈值选择的思想也体现了随机森林方法与认知诊断模型的结合。但是我们也应该注意到，这一方法是一种插补率和正确率之间的有效平衡，插补后的数据集仍存在少量的缺失数据。实际中，由于这一比例较小，可以将其视为可忽略的缺失值(Little & Rubin, 2002; Muthén et al., 2011)，以降低插补方法带来的不确定性。

研究 2 的模拟研究结果验证了在被试属性模式判准率上，RFTI 方法的有效性，以及与其他方法相比所表现出来的优势。与我们的预期一致，由于 RFTI 是一种非参数的缺失数据插补方法，其表现出较少受到缺失机制和缺失比例影响的优点。同时，由于其在缺失数据插补过程中，主要利用被试个体内的反应模式对其缺失的类别做出概率判断，可以充分利用 MIXED 和 MNAR 缺失机制下，模式反应上的差异提供的信息，因此，表现出在缺失机制为 MIXED 和 MNAR 时在被试属性掌握模式上有更为明显的优势。这一受缺失机制影响的模式与传统 IRT 模型并不一致，究其原因可能与认知诊断模型中对被试知识状态的估计实际上是掌握和不掌握的分类预测，而非连续的能力估计。以往研究也发现，缺失数据处理方法的性能与缺失机制有关，其

关系取决于具体的研究背景，包括分析模型和数据类型(分类或连续)(Dai, 2017; Zhuchkova & Rotmistrov, 2021)。因此，可以推测在 CDM 和 IRT 之间，缺失数据机制对传统方法的影响可能不同。另一方面，RFTI 方法在对被试知识状态进行估计时表现出的优势可能是因为其在数据插补和阈值确定过程中，关注的是个人拟合指标而非参数拟合指标，同时也可能这一插补过程更多地利用了个体反应模式的信息，MNAR 和 MIXED 的缺失机制相比于 MCAR 和 MAR 机制，其本身反而提供了一些额外有用的信息。但是，从项目参数的估计精度来看，RFTI 方法并没有表现出一致的明显优势。对于项目参数大部分条件下采用 EM 方法时的估计最精确。这可能与 RFTI 方法本身在训练模型的过程中本质上并不能有效利用同一项目不同被试个体的信息有关。

5.2 研究局限性与展望

本研究的重点是将机器学习的方法与认知诊断模型结合，对其可行性和效果进行了初步的检验，尚有许多值得进一步思考和研究的问题。

(1)本研究只考虑了 0-1 评分的情况，如何对方法改进进行多级评分的缺失数据的插补，应用于多级评分的认知诊断模型，还有待进一步的研究。(2)对于认知诊断模型的选择，虽然从理论上来讲，RFTI 适用于所有明确定义项目反应函数的认知诊断模型，但是本研究只结合 DINA 模型验证了基于个人拟合指数确定动态阈值插补方法的有效性，但是 RFTI 在其他认知诊断模型中，是否能够得到与本研究类似的结果，仍有待进一步验证。(3)由于本研究重点是探究缺失数据的处理，所以 RFTI 方法及对该方法优势的结论都是基于 Q 矩阵正确设定的前提，实际中 Q 矩阵的正确设定也是认知诊断测评关注的重要议题。当 Q 矩阵设定存在错误时，未来研究一方面可以探讨 RFTI 方法对缺失数据的插补效果如何受到 Q 矩阵错误设定以及错误设定程度的影响，另一方面在采用 RFTI 方法进行缺失数据插补前，可以对 Q 矩阵设定进行修正(Liu et al., 2021; 李佳 等, 2021)，基于修正后的 Q 矩阵再采用 RFTI 方法处理缺失值。(4)本研究没有对属性之间结构关系以及认知诊断模型中可能存在的项目特征相依(Zhan et al., 2019)带来的影响进行深入的探讨，未来的模拟研究可以设定更多的条件，进一步考查这些因素对 RFTI 方法可能产生的影响。(5)本研究确定阈值的过程是在指定范围内，按照事先定义的步长在区域内进行搜索，这一方法在实现虽然较为直接，但有可能效率较低，例如比较耗时；同时可能存在由于步长设置不同而使得结果存在细微的差异。在未来研究中，可以进一步探讨不同条件下，阈值变化与个人拟合指标的变化规律，在理论上推演二者的关系，为阈值的确定提供更充分的依据。

5.3 结论与建议

本研究得到的主要结论如下。

(1)本研究提出了一种 RFI 和 DINA 模型相结合的 RFTI 方法，该方法是一种不依赖于缺失机制假设的非参数插补方法。并开发了实现这一方法的 R 程序包，为实际应用者提供了方便易用的工具。

(2)RFTI 在正确率上弥补了 RFI 正确率过低的局限，并且对 DINA 模型的项目参数 s 和 g 参数的估计结果，以及被试属性掌握模式和属性边际判准率的估计结果均优于 RFI 方法。

(3)对于被试知识状态的估计结果表明，在考虑的所有条件下，RFTI 方法均优于 RFI 方法和 EM 方法，特别是在缺失机制为 MIXED 和 MNAR，以及缺失比例较高(≥30%)时，RFTI 方法的优势更加明显。

(4)项目参数估计结果表明在缺失比例较低或缺失机制为 MCAR 和 MNAR 时，EM 方法优于 RFI 和 RFTI 方法；在 MNAR 和 MIXED 缺失机制下，对于 *g* 参数的估计 RFTI 表现出优势。总体而言，采用 RFTI 方法在参数估计上的表现一般，与其他方法相比并不具备优势。

基于本研究的结论，我们给出 RFTI 方法选择上的建议如下：对于含有缺失数据的认知诊断，如果研究者关注的重点是被试知识状态的估计(这往往是认知诊断测验本身要解决的问题，是实际应用关注的焦点)，我们推荐使用新提出的 RFTI 方法；但是如果研究者的目的是对项目参数进行准确估计，如建立题库等，这一方法的使用则要相当慎重，我们则推荐采用 EM 算法。

参 考 文 献

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement Interdisciplinary Research & Perspectives, 8*(2–3), 70–91.

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research, 84*(4), 487–508.

Chen, P., & Xin, T. (2011). Item replenishing in cognitive diagnostic computerized adaptive testing. *Acta Psychologica Sinica, 43*(7), 836–850.

[陈平，辛涛．(2011)．认知诊断计算机化自适应测验中的项目增补．*心理学报*, *43*(7), 836–850.]

Chen Y., Li X., Liu J., & Ying Z. (2018). Recommendation system for adaptive learning. *Applied psychological measurement, 42*(1), 24–41.

Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: The modified maximum global discrimination index method. *Educational and Psychological Measurement, 70*(6), 902–913.

Cui, Y., & Li, L. (2015). Evaluating person fit for cognitive diagnostic assessment. *Applied Psychological Measurement, 39*(3), 223–238.

Dai, S. (2017). *Investigation of missing responses in implementation of cognitive diagnostic models.* (Unpublished doctoral dissertation). Indiana University, Indiana, Indiana U.

Dai, S., Svetina Valdivia, D. (2022). Dealing with missing responses in cognitive diagnostic modeling. *Psych*, 4, 318–342. https://doi.org/10.3390/psych4020028.

De Ayala, R. J., Plake, B. S. & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement, 38*(3), 213–234.

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement*, *47*(2), 227–249.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*(3), 225–245.

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*(2), 1–24.

Gierl, M. J., Wang, C., & Zhou, J. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT. *Journal of Technology, Learning, and Assessment*, *6*(6). Retrieved from http://www.jtla.org

Glas, C., & Pimentel, J. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 68*(6), 907–922.

Graham, J.W., Taylor, B.J., Olchowski, A.E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods. 11*,323–343.

Johnson, E. G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement*, *29*(2), 95–110.

Li, J., Mao, X., & Zhang, X. (2021). *Q*-matrix estimation (validation) methods for cognitive diagnosis. *Advances in Psychological Science, 29*(12), 2272–2280.

[李佳，毛秀珍，张雪琴. (2021)．认知诊断 Q 矩阵估计(修正)方法．*心理科学进展, 29*(12), 2272-2280.]

Little, R., & Rubin, D. B. (2002). *Statistical analysis with missing data: Second Edition*. New York: Wiley.

Liu, Y., Xin, T., & Jiang, Y. (2021). Structural parameter standard error estimation method in diagnostic classification models: Estimation and application. *Multivariate Behavioral Research, 57*(5), 784–803.

Liu, Y., Andersson, B., Xin, T., Zhang, H., & Wang, L. (2019). Improved Wald statistics for item-level model comparison in diagnostic classification models. *Applied Psychological Measurement, 43*, 402–414.

Liu,Y., Zhang, Q., Zheng, Z., & Yin, H. (2019). The Robustness of the item-level model comparison statistics in cognitive diagnostic models. *Journal of Psychological Science*, *42*(5), 1251–1259.

[刘彦楼，张倩萌，郑宗军，尹昊．(2019)．认知诊断模型中项目水平模型比较统计量的健壮性．*心理科学, 42*(5), 1251–1259.]

Liu, Y., & Gopalakrishnan, V. (2017). An overview and evaluation of recent machine learning imputation methods using cardiac imaging data. *Data, 2*(1), 8–23.

Luo, Z. S., Li, Y, J., Yu, X. F., Gao, C. L., & Peng, Y. F. (2015). A simple cognitive diagnosis method based on *Q*-Matrix theory. *Acta Psychologica Sinica*, *47*(2), 264–272.

[罗照盛，李喻骏，喻晓锋，高椿雷，彭亚风. (2015)．一种基于 Q 矩阵理论朴素的认知诊断方法．*心理学报, 47*(2), 264–272.]

Mabrey, D. J. (2006). *Tactical terrorism analysis: A comparative study of statistical learning techniques to predict culpability for terrorist bombings in two regionals low–intensity conflicts*. Unpublished doctoral Dissertation, Sam Houston State University, Huntsville, TX.

McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research*, *29*, 409–454.

Mislevy, R. J., & Wu, P. K. (1988). *Inferring examinee ability when some item responses missing* (RR–88–48–ONR). Princeton. NJ: Educational Testing Service.

Muthén, B., Asparouhov, T., Hunter, A., & Leuchter, A. (2011). Growth modeling with non-ignorable dropout: Alternative analyses of the STAR*D antidepressant trial. *Psychological Methods, 16*(1), 17−33.

Pan, Y., & Zhan, P. (2020). The impact of sample attrition on longitudinal learning diagnosis: A Prolog. *Frontiers in Psychology, 11*, 1051.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*(4), 525–556.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement, 74*(3), 423–452.

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory* (IRT) (ETS Research Rep. no. RR–10–11), Princeton, NJ: Educational Testing Service.

Schafer, J., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177.

Shen, L., Hu, G. Q., Chen, L. Z., & Tan, H. Z. (2014). Application of missforest algorithm for imputing missing data. *Chinese Journal of Health Statistics, 31*(5), 774–776.

[沈琳，胡国清，陈立章，谭红专．(2014)．缺失森林算法在缺失值插补中的应用．*中国卫生统计, 31*(5), 774–776.]

Song, Z. L., Guo, L., & Zheng, T. P. (2022). Comparison of missing data handling methods in cognitive diagnosis: Zero replacement, multiple imputation, and maximum likelihood estimation. *Acta Psychologica Sinica*, *54*(4), 426–440.

[宋枝璘，郭磊，郑天鹏．(2022)．认知诊断缺失数据处理方法的比较：零替换、多重插补与极大似然估计法．*心理学报, 54*(4), 426–440.]

Stekhoven, D. (2013). MissForest: Nonparametric missing value imputation using random forest. *R package version 1.4.*

Stekhoven, D., & Bühlmann, P. (2012). MissForest – nonparametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112–118.

Wang, P. J., Liu, H. Y. (2019). Make adaptive testing know examinees better: The item selection strategies based on

recommender systems. *Acta Psychologica Sinica*, *51*(9), 1057–1067.
[王璞珏, 刘红云. (2019). 让自适应测验更知人善选——基于推荐系统的选题策略. *心理学报*, *51*(9), 1057–1067.]
Wang, W. Y. (2012). *Researches on methods for aiding item attributes identifying in cognitive diagnostic assessment* (Unpublished doctoral dissertation. Jiangxi Normal University, China.
[汪文义. (2012). *认知诊断评估中项目属性辅助标定方法研究* (博士论文). 江西师范大学.]
Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika*. *81*(3), 625–649.
Yu, X. F., Luo, Z. S., Gao, C. L., & Qin, C. Y. (2014). Compare the diagnostic assessment classification accuracy when the Q-Matrix contains error. *Journal of Psychological Science*. *37*(6), 1482–1488.
[喻晓锋, 罗照盛, 高椿雷, 秦春影. (2014). Q 矩阵包含错误的认知诊断测验分类准确性研究. *心理科学*, *37*(6), 1482–1488.]
Zhan, P., Jiao, H., Liao, M., & Bian,Y. (2019). Bayesian DINA modeling incorporating within-item characteristic dependency. *Applied Psychological Measurement, 43*(2), 143–158.
Zhang S., & Chang, H. H. (2016). From smart testing to smart learning: How testing technology can assist the new generation of education. *International Journal of Smart Technology and Learning*, *1*(1), 67–92.
Zhuchkova, S., & Rotmistrov, A. (2021). How to choose an approach to handling missing categorical data: (un)expected findings from a simulated statistical experiment. *Quality & Quantity, 56*, 1–22. https://doi.org/10.1007/s11135-021-01114-w

Missing data analysis in cognitive diagnostic models: Random forest threshold imputation method

YOU Xiaofeng[1], YANG Jianqin[1], Qin Chunying[1], LIU Hongyun[2,3]
([1] *School of Mathematics and Information Science, Nanchang Normal University, Nanchang* 330022, *China*)
([2] *Beijing Key Laboratory of Applied Experimental Psychology, Beijing Normal University, Beijing* 100875, *China*)
([3] *Faculty of Psychology, Beijing Normal University, Beijing* 100875, *China*)

Abstract

In recent years, interest in cognitive diagnostic assessments (CDAs), as a new form of test, has increased drastically. Due to the specific design of the test, missing data is an inevitable problem in CDAs. Proper handling of missing data in CDAs is important to provide accurate diagnostic feedback to students and teachers. With the use of machine learning in education, relevant advancements have been made in missing data imputation. Research showed machine learning techniques have more desirable features for missing data imputation than traditional approaches. The random forest algorithm has been extended to become the random forest imputation (RFI) method in handling of CDAs missing data for CDAs. The method takes into consideration the characteristics of the data rather than assumes certain missing mechanism. RFI is a new non-parametric method that makes full use of the available response information and characteristics of response patterns to impute missing data.

Making use of advantages of RFI in categorization/prediction and its non-reliant on missing mechanism type, we improved and proposed the new random forest threshold imputation (RFTI) method. It could be used to impute missing responses in the widely used DINA (Deterministic Inputs, Noise “And” Gate) model. This research proposed to apply the Response Conformity Index (RCI) in the missing data imputation to set the threshold of imputation and to develop a method for missing response treatment for CDAs without totally relying on imputation. Two simulation studies were conducted to compare the performance of the proposed method and traditional models. Study 1 began by introducing the theoretical background and algorithm implementation of RFTI. Then, RFTI and RFI were compared in terms of accuracy rate of imputation for data with different proportions of missingness (10%, 20%, 30%, 40%, 50%) and missing data mechanisms (MIXED, MNAR, MAR, MCAR). This was to affirm the necessity of including RCI during imputation. Study 2 aimed to investigate the performance of RFTI, as well as RFI and EM algorithm in imputing missing data under different conditions. The manipulated design factors were identical to those in Study 1. We evaluated RFTI in terms of its accuracy in assessing the model attributes and item parameters. We also compared RFTI against the traditionally better performed EM and RFI under various design conditions to explore the advantages and conditions of using RFTI.

Results of Study 1 showed that RFTI, as compared to RFI, improved accuracy when imputation threshold was one. In various design conditions, RFTI imputation rate and accuracy were also better. Study 2 showed that RFTI outperformed other methods (RFI, EM algorithm) in accurately assessing the attribute pattern and attribute margin. This advantage was affected by the missing data mechanism and the proportion of missing data. Notably, RFTI was particularly better than other methods in handling mixed type of missing or MNAR data, and when the proportion of missing data was higher than 30%. However, RFTI was not any better than other methods in its accuracy of item parameter estimates. In most conditions, EM algorithm provided the most accurate parameter estimates.

In sum, we propose a method to impute missing data in CDAs by applying machine learning methods in measurement models. The advantage of this new method is affirmed through its accurate assessment of attribute pattern and attribute margin of DINA model. Theoretically, the current study provides a missing data imputation approach with less assumptions, which extends the traditional methods to impute missing data in CDAs framework. Moreover, we investigate how to estimate the attribute pattern of students accurately through the responses of a few items. It sheds lights on imputing missing data due to particularly designs in assessment or teaching.

Keywords missing data, cognitive diagnostic assessment, random forest threshold imputation, random forest imputation, expectation-maximization algorithm